Εργασία 1

Εισαγωγή

Η εργασία αυτή έχει ως στόχο την εμπέδωση των προβλημάτων του διαφορετικού κόστους εσφαλμένης ταξινόμησης και της ασυμμετρίας κλάσεων καθώς και των λύσεων που υπάρχουν σε αυτά τα προβλήματα.

Μέρος Α

Στο πρώτο μέρος της εργασίας θα δουλέψετε στο πρόβλημα του διαφορετικού κόστους ταξινόμησης. Θα χρησιμοποιήσετε ένα σύνολο δεδομένων για αξιολόγηση ρίσκου δανειοδότησης και τον πίνακα κόστους που το συνοδεύει[1]. Εφαρμόστε τρεις από τις τεχνικές που συζητήσαμε στο μάθημα ή παρεμφερείς. Μία τύπου sampling, μία τύπου weighting και μία τύπου minimizing expected cost. Συνδυάστε τες με τους αλγορίθμους μάθησης Random Forest, Linear SVM και Naive Bayes. Παρουσιάστε και σχολιάστε τα αποτελέσματα. Αυτό που θέλουμε να δούμε είναι αν οι τεχνικές βοηθάνε στο πρόβλημα και ποιες τα πηγαίνουν καλύτερα για κάθε αλγόριθμο μάθησης και συνολικά.

Μέρος Β

Στο δεύτερο μέρος της εργασίας θα δουλέψετε στο πρόβλημα της ασυμμετρίας κλάσεων χρησιμοποιώντας την βιβλιοθήκη imbalanced-learn. Θα χρησιμοποιήσετε ένα σύνολο δεδομένων για ανίχνευση απάτης στο πεδίο της ασφάλειας αυτοκινήτων[2]. Εφαρμόστε τρείς από τις τεχνικές που συζητήσαμε στο μάθημα ή παρεμφερείς. Μία τύπου υπερδειγματοληψίας με κατασκευή συνθετικών δεδομένων, μία τύπου υποδειγματοληψίας είτε με ομαδοποίηση είτε με καθαρισμό δεδομένων, και μία που είτε συνδυάζει υπέρ και υπόδειγματοληψία είτε χρησιμοποιεί κάποιο σύνολο μοντέλων (π.χ. EasyEnsemble). Συνδυάστε τες με τους αλγορίθμους μάθησης Random Forest, Linear SVM και Naive Bayes. Παρουσιάστε και σχολιάστε τα αποτελέσματα. Αυτό που θέλουμε να δούμε είναι αν οι τεχνικές βοηθάνε στο πρόβλημα και ποιες τα πηγαίνουν καλύτερα για κάθε αλγόριθμο μάθησης και συνολικά.

Λογιστικά

Οι εργασίες μπορούν να γίνουν είτε ατομικά, είτε σε ομάδες 2 ατόμων. Θα πρέπει να παραδώσετε από ένα Colab notebook για κάθε μέρος της εργασίας. Κάθε notebook θα περιέχει τον κώδικα που χρησιμοποιήσατε, θα παρουσιάζει τα αποτελέσματα των πειραμάτων σας και θα τα σχολιάζει. Θα πρέπει να παραδώσετε μέσω του elearning ένα αρχείο zip με τα δυο notebooks, αφού τα κατεβάσετε από το Colab ως ipynb αρχεία. Προθεσμία υποβολής: 16/4/2023.

---

[1] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[2] https://www.kaggle.com/datasets/khusheekapoor/vehicle-insurance-fraud-detection